

BEST AVAILABLE COPY

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-331490

(43)Date of publication of application : 30.11.2001

(51)Int.Cl.

G06F 17/30

G06F 12/00

G06F 17/21

(21)Application number : 2001-062753

(71)Applicant : FUJITSU LTD

(22)Date of filing : 07.03.2001

(72)Inventor : NITTA KIYOSHI
KOZAKURA FUMIHIKO

(30)Priority

Priority number : 2000075678

Priority date : 17.03.2000

Priority country : JP

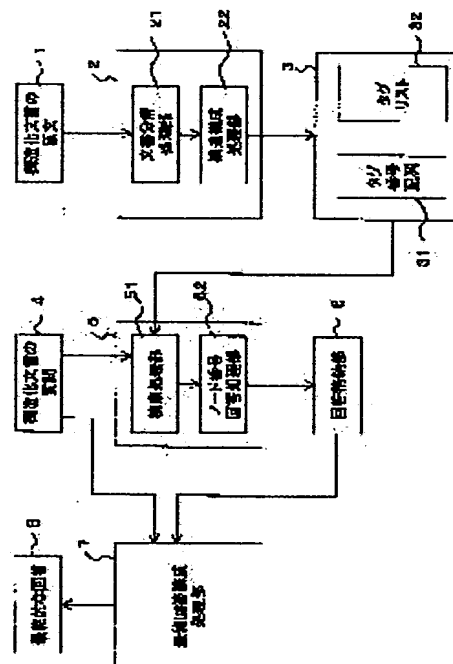
(54) STRUCTURED DOCUMENT STORAGE DEVICE, STRUCTURED DOCUMENT RETRIEVAL DEVICE, STRUCTURED DOCUMENT STORAGE AND RETRIEVAL DEVICE AND PROGRAM AND PROGRAM RECORDING MEDIUM

(57)Abstract:

PROBLEM TO BE SOLVED: To store a structured document suitable for retrieval at a high speed and the improvement of the using efficiency of memory resources regarding a structured document storage device.

SOLUTION: This structured document storage device is provided with a document decomposition processing part 21 for decomposing the original text 1 of the structured document into a tree structure composed of nodes and elements, a structure constitution processing part 22 for constituting a tag list structure based on the result of decomposition and a structure storage part 3 for storing at least the tag list structure. The tag list structure is composed of a tag number array 31 and a tag list 32.

The tag number array 31 stores tag numbers uniquely determined for the respective kinds of tags which are information for indicating the kinds of the elements for the tags of the elements included in the original text 1 of the structured document and pointers to the tag list 32 corresponding to them. The tag list 32 is composed by connecting the nodes provided with the tags of the same kind or the tags of the different kinds in mutually same relation.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-331490

(P2001-331490A)

(43) 公開日 平成13年11月30日 (2001. 11. 30)

(51) Int.Cl. ⁷	識別記号	F I	テームト [*] (参考)
G 0 6 F 17/30	1 4 0	G 0 6 F 17/30	1 4 0
	1 7 0		1 7 0 A
	2 3 0		2 3 0 Z
12/00	5 1 3	12/00	5 1 3 Z
17/21	5 0 1	17/21	5 0 1 T
審査請求 未請求 請求項の数16 O L (全 17 頁)			

(21) 出願番号 特願2001-62753(P2001-62753)

(22) 出願日 平成13年3月7日 (2001. 3. 7)

(31) 優先権主張番号 特願2000-75678(P2000-75678)

(32) 優先日 平成12年3月17日 (2000. 3. 17)

(33) 優先権主張国 日本 (J P)

(71) 出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中4丁目1番
1号

(72) 発明者 新田 清

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(72) 発明者 小櫻 文彦

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(74) 代理人 100074848

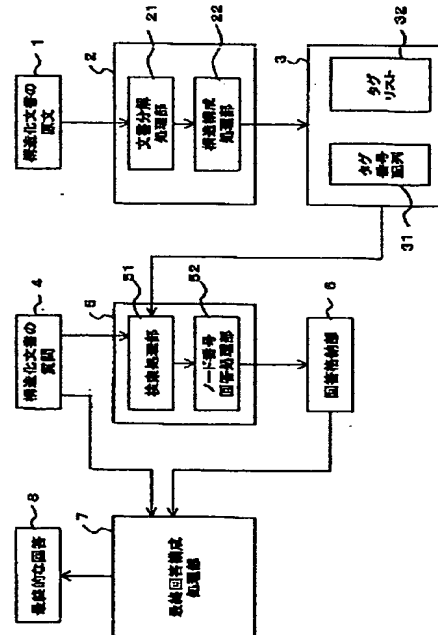
弁理士 森田 寛 (外1名)

(54) 【発明の名称】 構造化文書格納装置、構造化文書検索装置、構造化文書格納検索装置及びプログラム並びにプロ

(57) 【要約】 グラム記録媒体

【課題】 本発明は、構造化文書格納装置に関し、高速での検索及びメモリ資源の使用効率の向上に適した構造化文書を格納することを目的とする。

【解決手段】 構造化文書格納装置は、構造化文書の原文1をノード及び要素からなる木構造に分解する文書分解処理部21と、分解の結果に基づいて、タグリスト構造を構成する構造構成処理部22と、少なくともタグリスト構造を格納する構造格納部3とを備える。タグリスト構造はタグ番号配列31とタグリスト32とからなる。タグ番号配列31は、要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文1に含まれる要素のタグについてのタグ番号とこれに対応するタグリスト32へのポイントとを格納してなる。タグリスト32は、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つノードを連結してなる。



【特許請求の範囲】

【請求項1】 構造化文書の原文をノード及び要素からなる木構造に分解する文書分解処理部と、
前記分解の結果に基づいて、タグリスト構造を構成する構造構成処理部と、

少なくとも前記タグリスト構造を格納する構造格納部とを備え、

前記タグリスト構造はタグ番号配列とタグリストとからなり、

前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号とこれに対応する前記タグリストへのポイントとを格納してなり、

前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなることを特徴とする構造化文書格納装置。

【請求項2】 前記構造構成処理部が、当該構造化文書の原文から得られる前記ノード毎に、ノード番号を一意に定め、

前記タグリストにおいて、前記ノード番号を用いて前記ノードを連結することを特徴とする請求項1に記載の構造化文書格納装置。

【請求項3】 前記構造構成処理部が、当該構造化文書の原文から得られる前記要素毎に、要素番号を一意に定めることを特徴とする請求項2に記載の構造化文書格納装置。

【請求項4】 当該構造化文書格納装置が、更に、前記構造化文書の原文を格納する原文格納部を備えることを特徴とする請求項1に記載の構造化文書格納装置。

【請求項5】 当該構造化文書格納装置が、更に、前記構造化文書の原文を走査して、当該構造化文書の原文に含まれる要素についてのタグの種類総数を算出するタグ算出処理部を備えることを特徴とする請求項1に記載の構造化文書格納装置。

【請求項6】 構造化文書の原文をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、前記構造化文書についての質問に基づいて検索する検索処理部を備え、

前記タグリスト構造はタグ番号配列とタグリストとからなり、

前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号及びこれに対応する前記タグリストへのポイントとを格納してなり、

前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなり、

前記検索処理部が、前記質問の内容に該当するタグにつ

いてのタグ番号を用いて、前記タグ番号配列から当該タグ番号に対応する前記タグリストへのポイントを参照し、当該ポイントを用いて前記タグリストに連結された前記ノードを検索することを特徴とする構造化文書検索装置。

【請求項7】 前記タグリストにおいて、前記構造化文書の原文から得られる前記ノード毎に一意に定められるノード番号を用いて、前記ノードが連結され、前記検索処理部が、当該検索の結果として、当該ノードのノード番号を得ることを特徴とする請求項6に記載の構造化文書検索装置。

【請求項8】 構造化文書の原文をノード及び要素に分解した木構造を用いて、前記構造化文書についての質問に対する回答を得るノード番号回答処理部を備え、

前記構造化文書の原文から得られる前記ノード毎にノード番号が一意に定められ、

前記ノード番号回答処理部が、前記質問に対する回答を、前記質問の内容に該当するノードのノード番号、又は、前記木構造において前記質問の内容に該当する範囲を規定するノードのノード番号で構成することを特徴とする構造化文書検索装置。

【請求項9】 当該構造化文書検索装置が、更に、前記構造化文書の原文の文書名、前記質問の内容、及び、前記ノード番号で構成された組を、最終回答として構成する最終回答構成処理部を備えることを特徴とする請求項8に記載の構造化文書検索装置。

【請求項10】 構造化文書の原文をノード及び要素からなる木構造に分解する文書分解処理部と、
前記分解の結果に基づいて、タグリスト構造を構成する構造構成処理部と、

少なくとも前記タグリスト構造を格納する構造格納部と、

構造化文書の原文をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、前記構造化文書についての質問に基づいて検索する検索処理部とを備え、
前記タグリスト構造はタグ番号配列とタグリストとからなり、

前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号とこれに対応する前記タグリストへのポイントとを格納してなり、

前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなり、

前記検索処理部が、前記質問の内容に該当するタグについてのタグ番号を用いて、前記タグ番号配列から当該タグ番号に対応する前記タグリストへのポイントを参照し、当該ポイントを用いて前記タグリストに連結された前記ノードを検索することを特徴とする構造化文書格納

検索装置。

【請求項 1 1】 コンピュータに実行させることにより構造化文書格納装置を実現するプログラムであって、構造化文書の原文をノード及び要素からなる木構造に分解する文書分解処理と、前記分解の結果に基づいて、タグリスト構造を構成する構造構成処理とを前記コンピュータに実行させ、前記タグリスト構造はタグ番号配列とタグリストとからなり、前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号とこれに対応する前記タグリストへのポイントとを格納してなり、前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなることを特徴とするプログラム。

【請求項 1 2】 コンピュータに実行させることにより構造化文書格納装置を実現するプログラムを格納するコンピュータ読取可能なプログラム記録媒体であって、前記プログラムは、構造化文書の原文をノード及び要素からなる木構造に分解する文書分解処理と、前記分解の結果に基づいて、タグリスト構造を構成する構造構成処理とを前記コンピュータに実行させ、前記タグリスト構造はタグ番号配列とタグリストとからなり、前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号とこれに対応する前記タグリストへのポイントとを格納してなり、前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなることを特徴とするプログラム記録媒体。

【請求項 1 3】 コンピュータに実行させることにより構造化文書検索装置を実現するプログラムであって、構造化文書の原文をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、前記構造化文書についての質問に基づいて検索する検索処理を前記コンピュータに実行させ、前記タグリスト構造はタグ番号配列とタグリストとからなり、前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号及びこれに対応する前記タグリストへのポイントとを格納してなり、前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結して

なり、

前記検索処理が、前記質問の内容に該当するタグについてのタグ番号を用いて、前記タグ番号配列から当該タグ番号に対応する前記タグリストへのポイント参照し、当該ポイントを用いて前記タグリストに連結された前記ノードを検索する処理であることを特徴とするプログラム。

【請求項 1 4】 コンピュータに実行させることにより構造化文書検索装置を実現するプログラムを格納するコンピュータ読取可能なプログラム記録媒体であって、前記プログラムは、構造化文書の原文をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、前記構造化文書についての質問に基づいて検索する検索処理を前記コンピュータに実行させ、前記タグリスト構造はタグ番号配列とタグリストとからなり、

前記タグ番号配列は、前記要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文に含まれる要素のタグについてのタグ番号及びこれに対応する前記タグリストへのポイントとを格納してなり、

前記タグリストは、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つ前記ノードを連結してなり、

前記検索処理が、前記質問の内容に該当するタグについてのタグ番号を用いて、前記タグ番号配列から当該タグ番号に対応する前記タグリストへのポイント参照し、当該ポイントを用いて前記タグリストに連結された前記ノードを検索する処理であることを特徴とするプログラム記録媒体。

【請求項 1 5】 コンピュータに実行させることにより構造化文書検索装置を実現するプログラムであって、構造化文書の原文をノード及び要素に分解した木構造を用いて、前記構造化文書についての質問に対する回答を得るノード番号回答処理を前記コンピュータに実行させ、

前記構造化文書の原文から得られる前記ノード毎にノード番号が一意に定められ、

前記ノード番号回答処理が、前記質問に対する回答を、前記質問の内容に該当するノードのノード番号、又は、前記木構造において前記質問の内容に該当する範囲を規定するノードのノード番号で構成する処理であることを特徴とするプログラム。

【請求項 1 6】 コンピュータに実行させることにより構造化文書検索装置を実現するプログラムを格納するコンピュータ読取可能なプログラム記録媒体であって、前記プログラムは、構造化文書の原文をノード及び要素に分解した木構造を用いて、前記構造化文書についての質問に対する回答を

得るノード番号回答処理を前記コンピュータに実行させ、
前記構造化文書の原文から得られる前記ノード毎にノード番号が一意に定められ、
前記ノード番号回答処理が、前記質問に対する回答を、前記質問の内容に該当するノードのノード番号、又は、前記木構造において前記質問の内容に該当する範囲を規定するノードのノード番号で構成する処理であることを特徴とするプログラム記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、構造化文書格納装置、構造化文書検索装置、構造化文書格納検索装置及びプログラム並びにプログラム記録媒体に関し、特に、非定型な構造化文書を高速で効率よく検索することができる構造化文書格納装置、構造化文書検索装置、構造化文書格納検索装置及びプログラム並びにプログラム記録媒体に関する。

【0002】

【従来の技術】企業等の組織においては、情報（データや知識）を当該組織内において流通、加工、共有することが求められる。このために、情報を共有することを目的とするシステム（ナレッジマネジメントシステム等）では、情報をXML等のタグ付き言語で表現した文書（構造化文書）として作成して蓄積（格納）し、これを検索し、当該検索結果を利用する。このような構造化文書の利用は、多くの場合、以下の3つの方式のいずれかによっている。

【0003】第1に、作成された構造化文書をそのまま格納し、その使用時に毎回、格納された構造化文書の全てを構造解析する方式がある。この方式は、例えば、特開平6-259421号公報「文書処理装置」等に開示されている。

【0004】第2に、構造化文書を予め構造解析し、その結果を表形式のデータベースに変換して格納する方式がある。この方式は、例えば、特開平5-225240号公報「文書データベース装置」、特開平6-119331号公報「構造化文書の文書部品管理装置」、特開平7-44579号公報「論理構造文書検索方式」、特開平7-319918号公報「文書検索対象指示装置」、特開平8-147311号公報「構造化文書検索方法及び装置」、特開平10-171794号公報「動的部品化機能付き構造化文書データベースシステム」等に開示されている。

【0005】第3に、構造化文書を予め構造解析し、その結果である「要素」と「構造情報」とを格納する方式がある。この方式は、例えば、特開平7-65035号公報「構造化文書検索装置」、特開平10-187680号公報「単語、文、部分の粒度で管理するドキュメントリポジトリ装置」等に開示されている。

【0006】

【発明が解決しようとする課題】前述した第1乃至第3の方式には以下のような技術的な問題があることが、一般に知られている。

【0007】即ち、第1の方式は、構造化文書の検索処理や回答（部分取り出し）処理等において格納された情報（構造化文書）を利用するが、各構造化文書をそのまま格納しているため、これについて利用の都度に構造解析を行ってから検索処理等を行なう。このため、検索処理等の効率が良くなく、処理速度も高速とは言えず、大規模なデータには適用が難しい。

【0008】第2の方式は、データベース（R（関係）DB、表DB等）の構成（フィールドの構成）を予め決定し、構造化文書のどの部分を当該データベースのどのフィールドに格納するかを設計し、その後に当該データベースへの格納処理を行なう。従って、個々に（相互に）構造が異なるような複数の構造化文書の格納及び検索には適していない。また、通常のRDBを用いる場合には、そのフィールドの最大サイズが決まっているので、解析結果に当該サイズ以上の部分を含む場合には、当該構造化文書のデータベースへの格納は難しい。

【0009】第3の方式は、構造化文書の格納時にその構造解析を行なって、要素と要素間の構造情報に分離し、その状態のまま格納を行なう。第3の方式によれば、予め構造解析が行われた状態で格納されているので、第1の方式よりは検索処理等の効率がよく大規模なデータにも適用できる。また、構造解析の結果をそのまま格納しているので、格納（しようと）する構造化文書の構造やデータサイズの制約を受けないため、第2の方式よりは格納の自由度が高く、適用できる構造化文書の範囲が制限されにくい。

【0010】しかし、第3の方式にも以下のような技術的な問題がある。第1に、第3の方式において格納される構造情報は、通常、木構造である。従って、検索処理において検索対象となる要素を取得するためには、「根（ルート）」と呼ばれる特定の要素から検索対象となる要素に至るまで、順に木構造をたどる必要がある。このため、検索速度が十分に高速であるとは言えない。また、木構造をたどる以外には検索対象となる要素を取り出す方法がなく、当該要素を直接取り出すことはできない。第2に、第3の方式において、構造情報は検索処理の手段にすぎず、検索処理の結果として得られるのは当該要素である。従って、検索処理の結果としては、当該結果である要素の実体を複製して蓄積しなければならず、当該結果の蓄積のためには多くのメモリ資源を消費する。このため、検索結果（複数の要素）を一時的に大量に蓄積して、これに対して（ソート等の）操作を行ってから、例えば上位100件等を選別して出力しようとする、膨大なメモリ資源を占有することになり、事実上困難である。

【0011】本発明は、高速での検索及びメモリ資源の使用効率の向上に適した構造化文書を格納する構造化文書格納装置を提供することを目的とする。

【0012】また、本発明は、構造化文書を高速で検索しその結果を効率よく格納する構造化文書検索装置を提供することを目的とする。

【0013】また、本発明は、高速での検索及びメモリ資源の使用効率の向上に適した構造化文書を格納する構造化文書格納装置を実現するプログラム、及び、当該プログラムを記録するプログラム記録媒体を提供することを目的とする。

【0014】また、本発明は、構造化文書を高速で検索しその結果を効率よく格納する構造化文書検索装置を実現するプログラム、及び、当該プログラム記録媒体を提供することを目的とする。

【0015】

【課題を解決するための手段】図1は本発明の原理構成図であり、本発明による構造化文書格納検索システムの構成を示す。

【0016】本発明の構造化文書格納装置は、構造化文書の原文1をノード及び要素からなる木構造に分解する文書分解処理部21と、分解の結果に基づいて、タグリスト構造を構成する構造構成処理部22と、少なくともタグリスト構造を格納する構造格納部3とを備える。タグリスト構造はタグ番号配列31とタグリスト32とからなる。タグ番号配列31は、要素の種類を表す情報であるタグの種類毎に一意に定められたタグ番号であって、当該構造化文書の原文1に含まれる要素のタグについてのタグ番号とこれに対応するタグリスト32へのポインタとを格納してなる。タグリスト32は、同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つノードを連結してなる。

【0017】本発明の構造化文書格納装置によれば、木構造に基づいて生成されたタグ番号配列31とタグリスト32とからなるタグリスト構造を得ることができる。このタグリスト構造は、後述するように、当該構造化文書についての質問4に基づく検索を高速に行うことができる構造である。従って、構造化文書の高速度での検索に適したデータ構造を得ることができる。

【0018】また、本発明の構造化文書検索装置は、構造化文書の原文1をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、構造化文書についての質問4に基づいて検索する検索処理部51を備える。タグリスト構造は、前述と同様の構成である。検索処理部51が、質問4の内容に該当するタグについてのタグ番号を用いて、タグ番号配列31から当該タグ番号に対応するタグリスト32へのポインタを参照し、当該ポインタを用いてタグリスト32に連結されたノードを検索する。

【0019】本発明の構造化文書検索装置によれば、木

構造に基づいて生成されたタグ番号配列31とタグリスト32とからなるタグリスト構造を利用して検索を行う。即ち、タグ番号配列31から構造化文書の原文1に含まれる要素のタグについてのタグ番号を求め、これを用いて対応するタグリスト32へのポインタを求め、当該ポインタにより同一種類のタグ又は相互に同一の関係である異なる種類のタグを持つノードを得る。従って、検索対象となる要素に至るまで順に木構造をたどることなく、かつ、検索対象となる要素以外の要素を全く検索することなく、検索対象となる要素を極めて高速に、検索対象となる要素を検索することができる。

【0020】また、本発明の構造化文書検索装置は、構造化文書の原文1をノード及び要素に分解した木構造を用いて、構造化文書についての質問4に対する回答を得るノード番号回答処理部52を備える。構造化文書の原文1から得られるノード毎にノード番号が一意に定められる。ノード番号回答処理部52が、質問4に対する回答を、質問4の内容に該当するノードのノード番号、又は、木構造において質問4の内容に該当する範囲を規定するノードのノード番号で構成する。

【0021】本発明の構造化文書検索装置によれば、質問4に対する回答が、一意に定められるノード番号により構成される。従って、検索処理の結果として当該結果である要素の実体を複製して蓄積する必要がなく、少ないメモリ資源で当該結果を蓄積することができる。このため、検索結果（複数の要素）を一時的に大量に蓄積して、これに対して（ソート等の）操作を行なって選別して出力するような処理を、容易に行うことができる。

【0022】更に、本発明の一実施態様によれば、当該構造化文書検索装置が、更に、構造化文書の原文の文書名、質問の内容、及び、ノード番号で構成された回答の組を、最終回答として構成する最終回答構成処理部7を備える。

【0023】これにより、質問4に対する回答が、短い文字列である文書名及び質問の内容と前述のノード番号とにより構成される。従って、検索処理の結果として当該結果である要素の実体を複製して蓄積する必要がなく、少ないメモリ資源で当該結果を蓄積することができる。

【0024】以上の処理をコンピュータによって実現するためのプログラムは、コンピュータが読取可能な種々の半導体メモリ、ディスクメモリ、その他の可搬メモリ等の記録媒体に格納される。

【0025】

【発明の実施の形態】図2は構造化文書格納検索システム構成図であり、本発明の構造化文書格納検索システムの構成を示す。構造化文書格納検索システムは、構造化文書格納装置である構造化文書格納処理部（処理モジュール、以下同じ）2、構造化文書検索装置の一部である構造化質問処理部5、構造化文書検索装置の一部である

最終回答構成処理部7からなる。これらの処理部は、1個のコンピュータ上で実現されてもよく、各々が独立したコンピュータ上で実現されてもよい。

【0026】構造化文書格納処理部2は、文書分解処理部21、構造構成処理部22、原文格納処理部25、構造格納部3、原文格納部9を備える。構造構成処理部22は、ノード番号付与処理部23、タグリスト構成処理部24を備える。

【0027】構造化文書格納処理部2は、与えられた（入力された）構造化文書の原文1を読み込む。原文格納処理部25は、当該読み込まれた原文1（図3参照）を、必要に応じて、原文格納部9に格納する。従って、原文格納部9は、構造化文書の原文1をそのまま格納する。文書分解処理部21は、当該読み込まれた構造化文書の原文1を分解処理して、当該原文1についてのノード及び要素からなる木構造33（図4参照）を得る。木構造33を得る手段は周知のいずれの手段によってもよい。

【0028】構造構成処理部22は、当該読み込まれた原文1の分解の結果である木構造33に基づいて、ノード番号付与処理及びタグリスト構造を構成する処理を行う。即ち、ノード番号付与処理部23が、木構造33のノードにノード番号を付与し、要素に要素番号を付与する。このために、構造構成処理部22又はノード番号付与処理部23は、当該構造化文書の原文1から得られるノード毎に、ノード番号を一意に定め、タグリスト32（図5参照）において、ノード番号を用いてノードを連結し、また、当該構造化文書の原文1から得られる要素毎に、要素番号を一意に定める。また、タグリスト構成処理部24が、タグ番号配列31（図5参照）とタグリスト32とからなるタグリスト構造（図5参照）を構成する。木構造33及びタグリスト構造は、構造格納部3に格納される。従って、構造格納部3は、検索処理のために、少なくともタグリスト構造を格納する。

【0029】なお、構造化文書格納処理部2が、タグ算出処理部（図示せず）を備えるようにしてもよい。タグ算出処理部は、文書分解処理に先立って、前述の読み込まれた構造化文書の原文1を走査して、当該構造化文書の原文1に含まれる要素についてのタグの種類総数を算出する。これにより、未知の構造化文書（例えば、文書型定義を参照することにより予め全てのタグを知ることができない文書）の原文1であっても、タグリスト構造の生成に先立って予めタグの総数を知り、後述するように、タグ番号を一意に定めることができる。

【0030】構造化質問処理部5は、検索処理部51、ノード番号回答処理部52、回答格納部6を備える。構造化質問処理部5は、与えられた（入力された）構造化文書の質問4を読み込み、当該質問4に対する回答（一時回答）を作成して、回答格納部6に格納する。即ち、検索処理部51が構造格納部3に格納されたタグリスト

構造を用いて当該質問4に基づく検索を行い、この検索の結果を用いて、ノード番号回答処理部52がノード番号を用いて表した回答（一時回答）を生成して、回答格納部6に格納する。

【0031】検索処理部51は、構造化文書の原文1をノード及び要素からなる木構造に分解して構成されたタグリスト構造を、構造化文書についての質問4に基づいて検索する。検索処理部51は、質問4の内容に該当するタグについてのタグ番号を用いて、タグ番号配列31から当該タグ番号に対応するタグリスト32へのポイントを参照し、当該ポイントを用いてタグリスト32に連結されたノードを検索する。タグリスト32において、構造化文書の原文1から得られるノード毎に一意に定められるノード番号を用いてノードが連結されているので、検索処理部51は、当該検索の結果として、当該ノードのノード番号を得る。

【0032】ノード番号回答処理部52は、構造化文書の原文1をノード及び要素に分解した木構造を用いて、構造化文書についての質問4に対する回答を得る。ノード番号回答処理部52は、質問4に対する回答を、質問4の内容に該当するノードのノード番号で構成する。又は、ノード番号回答処理部52は、質問4に対する回答を、木構造において質問4の内容に該当する範囲を規定するノードのノード番号で構成する。いずれの処理を行うかは、質問4の内容に依存する。

【0033】最終回答構成処理部7は、構造格納部3から木構造33及びタグリスト構造を読み出し、回答格納部6から一時回答を読み出し、構造化文書の質問4を読み込み、これらに基づいて最終的な回答8を生成して出力する。具体的には、最終回答構成処理部7は、構造化文書の原文1の文書名、質問4の内容、及び、ノード番号で構成された組を、最終回答8として構成する。

【0034】ここで、構造化文書の原文1、その木構造33、そのタグリスト構造について説明する。

【0035】図3は構造化文書の原文1の一例を示す。この文書の文書名（又はファイル名）は「0323. txt」であり、当該システムにおいて一意に定まり、これにより当該文書を識別できる。この原文1はXMLで記述されているが、SGML、HTML等のタグ付き言語で記述されている文書であればよい。即ち、構造化文書はタグ付き言語で記述された文書（タグ付き文書）である。図3において、例えば、<root>及び</root>等の「<>」で囲まれた部分がタグであり、「<>」及び「</>」が一对をなす。また、例えば、タグ<title>及び</title>に挟まれた文字列「AA社の株価」が要素である。

【0036】図4は図3に示す構造化文書の原文1についての木構造33を示す。即ち、当該原文1を要素に分解して得られる木構造33を示す。木構造33はノードと要素とを含む。ノード及び要素には、各々、ノード番

号及び要素番号が割り当てられる。図4において、角の丸い領域はタグに相当するノードを示す。ノードの横に記された数字はタグノードの通し番号を示す。ノード番号を用いた一時回答ではこの数字が用いられる。ノード番号には「\$」を付して表す。また、矩形領域はタグに囲まれた文書の要素を示す。これらの横に記された数字は要素の通し番号を示す。要素番号には「%」を付して表す。最上部の「root」タグに相当するノードが、当該木構造33のルートを示す。上下をつなぐ線分はノードどうしの親子関係（図中、上が親で下が子）を示す。なお、ノードの属性についても、ノード番号等と同様に、この木構造33に融合して格納してもよい。

【0037】従って、この明細書において、構造化文書の原文1を解析して分解した各要素を「ノード」、解析の結果構築する構造情報を「木構造」、ノードの種類を現わす情報を「タグ」（構造化文書表現におけるタグと同意）という。木構造33の一部分であって、分岐を含まない部分木を「分岐なし部分木」という。即ち、「分岐なし部分木」とは、当該部分木の全ての構成要素について子へのリンク数が高々「1」であるものをいう。木構造33における根（ルート）に相当するノードを「ルートノード」と呼ぶ。ルートからノードまでたどったときのタグのリストを「パス」と呼ぶ。構造化文書の重複しない名前又は識別子等を「文書のポイント」と呼び、その他のデータについてもそれを識別できる情報を同様に「ポイント」と呼ぶ。

【0038】図5は図3に示す構造化文書の原文1についてのタグリスト構造を示す。タグリスト構造は、当該原文1及びこれを要素に分解して得られる木構造33に基づいて得られる。タグリスト構造は、図5（A）に示すように、タグ番号配列31とタグリスト32（32A）とからなる。

【0039】タグ番号配列31は、当該構造化文書の原文1に含まれる要素のタグについてのタグ番号と、これに対応するタグリスト32へのポイントとを格納してなる。タグ番号は要素の種類を表す情報であるタグの種類毎に一意に定められる。具体的には、図5（B）に示すように、タグ番号配列31は、タグとそのタグ番号とを格納する表である。例えば、要素「株価」の種類を表す情報であるタグ<株価>に対して、タグ番号「#80」が一意に定まる。タグ番号には#を付して表す。タグ番号#80には、これに対応するタグリスト32へのポイントP1が付加される。タグ番号はタグ番号配列31のインデックスとなる。タグ番号インデックスに対応する文字列がタグ名である。

【0040】タグ番号及び要素番号は、例えば次のように定められる。木構造33の生成時において、各要素及びノードの原文1からの分離の順序をインデックスとし、ノードのポイントと値とする配列構造として生成する。当該配列におけるインデックス値が当該ノードのタ

グの通し番号となる。要素についても同様である。

【0041】従って、図4及び図5（A）に示すように、図3の文書1の先頭の2行（XMLでのきまりの部分）を無視して、次の<root>から順にノード番号が与えられる。要素についても、最初に現われる「AA社の株価」から順に要素番号が与えられる。

【0042】タグリスト32は、同一種類のタグを持つノードを連結してなる。即ち、タグ<株価>を持つノードを連結している。図3及び図4において、タグ<株価>を持つノードはノード番号\$8のノード及びノード番号\$11のノードであるので、ポイントP1及びP2により、これが連結される。タグ番号#80に付加されたポイントP1は、ノード番号\$8のノードを示す。ノード番号\$8のノードに付加されたポイントP2は、ノード番号\$11のノードを示す。ノード番号\$11のノードに付加されたポイントは、連結するノードがないので、空とされる。これらのポイントP1及びP2は、実際は、当該ノード自体ではなく、そのノード番号（の格納されている位置）を示す。

【0043】なお、タグリスト32は、同一種類のタグのみでなく、相互に同一の関係である異なる種類のタグを持つノードを連結してなるようにしてもよい。例えば、タグ番号#79であるタグ<会社名>は、<企業名>又は<株式会社名>等であっても同義である。従って、異なる種類のタグであっても、相互に同一の関係であるタグを持つノードを連結して、タグリスト32を生成してもよい。

【0044】構造化文書格納検索システムにおいては、以上の構成の構造化文書の原文1、木構造33及びタグリスト構造（31、32）の格納処理と、これらを用いて構造化文書について質問4に基づく検索・回答処理が行われる。検索・回答処理としては、一時回答処理と最終回答処理とが行われる。

【0045】格納処理においては、図6に示すように、格納する構造化文書の原文1を取得した上で、必要に応じて、複数の当該原文1A乃至1Cの格納を行なう。原文1Aの格納されたものを原文92Aと表す。原文1Aと原文92Aとは同一である。複数の構造化文書の原文92A乃至92Cが格納され、格納された文書群91を構成する。文書分解処理により原文92A乃至92Cの各々を、分解（タグで区切られる要素をそれぞれに分解）し、各々の木構造33A乃至33Cを構成し、ノード番号及び要素番号を付与し、タグリスト構造を構成する。タグリスト構造において、タグ番号配列31は文書群91に対応して1個生成される。従って、タグ番号配列31は原文92A乃至92Cに共通であり、タグ番号は当該文書群91において一意に定まる。一方、タグリスト32A乃至32Cは、原文92A乃至92Cの各々に対応して個々に生成される。

【0046】なお、当該システムにおいて、文書群91

を複数（図示せず）格納するようにしてもよい。この場合でも、図6に示すように、文書群91毎にタグ番号配列31が生成される。そして、1個の文書群91及びタグ番号配列31につき、これに含まれる文書1（92）の数だけタグリスト32が生成される。

【0047】構造格納部3には、1個のタグ番号配列31、タグリスト32A乃至32C、木構造33A乃至33C、ノード及び要素34A乃至34Cが格納される。タグ番号配列31、タグリスト32A乃至32C、木構造33A乃至33Cが全てノード番号、ポイント等それ自体では意味のない情報からなるのに対して、ノード及び要素34A乃至34Cは、各々、原文92A乃至92Cの実際の要素及びノードそのものである。例えば、ノード及び要素34Aの前半部分には対応する原文92Aのノードが格納され、対応する木構造33Aの対応するノード番号によりポイントされ、後半部分には対応する原文92Aの要素が格納され、対応する木構造33Aの対応する要素番号によりポイントされる。

【0048】検索・回答処理においては、図7に示すように、構造化文書の質問4を取得して、これに基づいて検索処理部51が検索処理を行う。これにより当該質問の内容（条件）に該当する文書及びノード（ノード番号）を得る。この検索処理において、タグリスト構造（31、32）及び木構造33を利用して、検索時における無駄な検索処理を抑制する。これにより、検索の高速化を図っている。

【0049】即ち、どのような手法の構造検索を行なう場合でも、文書中に含まれる特定のタグ番号を持つ文書要素（ノード）を探す処理が行なわれる。タグリスト構造がない場合、その検索処理において、その都度ルートノードから文書全体をたどりつくす必要がある。これに対して、タグリスト構造は、その格納時に1回（タグ種類数が未知の場合は2回）木構造33をたどることで生成できる。一旦生成してしまえば、検索時には文書全体をたどらずに、タグリスト構造にアクセスするだけで検索処理を行うことができる。これにより、ノード検索における実行ステップ数を削減し、検索時間を短縮することができる。なお、データの形態や検索条件によってその効果の程度は変化する。また、本発明のタグリスト構造による検索は、タグ番号からノードを検索する処理を短縮させるので、ノード数が多い文書が多数を占めるデータで、葉に近いノードを回答指示することが多い場合に比較的效果が大きい。

【0050】ここで、構造化文書の質問4について説明する。図8（A）は構造化文書の質問4の一例を示す。

【0051】この質問4は、SELECT部、WHERE部、ORDERBY部からなる。この質問4は、タグ<会社情報>のすぐ内側にタグ<会社名>があり、タグ<会社名>で回された内容が<AA社>であるような文書について（WHERE部）、条件を満たすタグ<会社

情報>直下のタグ<株価>の内容を選択し（SELECT部）、その結果をその株価でソートせよ（ORDERBY部）、という指示を意味する。即ち、「AA社の株価を、その高い順に出力せよ」という回答指示である。なお、質問4において、SELECT部及びORDERBY部の「#i」は、SELECT部及びORDERBY部がWHERE部の「#i」に連動するとを示す。即ち、SELECT部及びORDERBY部には<AA社>の記述がないが、<AA社>についての処理であることを指示している。

【0052】SELECT部の内容「株価」が回答指示要素sである。回答指示要素は、s1、s2、・・・のように複数記述することができる。回答指示要素s1、s2、・・・が複数の場合、図8（B）に示すように、各々の要素毎に、回答（一時回答）の集合が得られる。回答の集合は、文書毎ではなく、要素毎に得られる。例えば、要素s1について、回答の集合（dp, s1, Nn(d, s)）が得られる。各記号については後述する。

【0053】図7において、このような質問4を取得した検索処理部51は、SELECT部の「株価」を用いてタグ番号配列31をアクセスして、図5（B）に示すように、そのタグ番号#80を知り、これを用いて文書群91の全ての文書、即ち、全てのタグリスト32を検索する。これにより、「株価」が文中に現れる文書及びそのノード（実際はノード番号及びそのポイント）を、検索結果60として得る。例えば、図3に示す文書の文書名「0323.txt」と、これにおける「株価」の現れるノード番号\$8及び\$11を得る。この時点では、ノード番号\$11はBB社の株価であるので、回答としては採用できない。この検索結果60は、例えば回答格納部6に一時的に格納される。従って、この検索において、関連のない文書及びノードを検索することなく、極めて高速に関連する文書名及びノード番号を得ることができる。

【0054】このような検索結果60を取得したノード番号回答処理部52は、これに基づいて、第1ノード番号回答処理部521において質問4に対する回答61を生成するか、又は、第2ノード番号回答処理部522において質問4に対する回答62を生成する。回答61は、質問4の内容に該当するノードのノード番号で構成する。例えば、ノード番号\$8である。回答62は、木構造において質問4の内容に該当する範囲を規定するノードのノード番号で構成する。例えば、回答指示が<root><article><date>である場合、上位のノード番号\$1及び下位のノード番号\$4である。

【0055】このようにノード番号を用いた一時回答処理により回答格納部6に一時的な回答を格納し、最終回答処理により一時結果をソート等により選別し、ノード

番号の参照先は構造格納部3から得ることにより最終回答8を構成する。従って、ノード番号を格納し、検索回答にはそのノード番号を用いて一時的な結果を蓄積する。これにより、一時結果の蓄積時のメモリサイズを抑制することができる。一般に、構造検索におけるひとつの回答指定は原文1の一部（分岐なし部分木）となる。一時結果をノード番号を用いずに蓄積した場合、一時結果の件数及び文書サイズに比例する記憶領域が必要になる。しかし、一時結果をノード番号の形で蓄積することにより、メモリサイズを小さくでき、また文書サイズに影響を受けないようにできる。なお、一時回答を選別せずに全て出力する場合、ソート処理及び最終回答構成処理は省略してよい。

【0056】その構成の手段が回答61又は62のいずれであっても、回答指示に対する直接の回答となるノードは、別に定まる。例えば、回答指示が「AA社の株価」であれば、前述の例において、ノード番号\$8のみが選択され、BB社の株価であるノード番号\$11は選択されない。即ち、前述したように、WHERE部に従って、タグ<会社情報>のすぐ内側にタグ<会社名>があり、タグ<会社名>で回まれた内容が<AA社>であるような文書について（従って、ノード番号\$11は除かれる）、条件を満たすタグ<会社情報>直下のタグ<株価>の内容が選択される。この結果、例えば、図9

(A)に示すような結果が、回答61又は62として得られる。図9(A)において、一行が一件の回答を示す。各回答は、ヒットした文書の文書名（を示すポインタ）dp、質問4の回答指示要素（SELECT部の指示）s1、文書dp内のノード番号Nn（d，s）、ソート用の値からなる。

【0057】回答61又は62は、操作処理部10において、必要に応じて、ソート等の操作を受ける。なお、ソート等の処理は最終回答構成処理の後に行ってもよい。操作処理部10は前述のORDERBY部を実行する。操作処理部10は、図7に点線で示すように、当該構造化文書格納検索システム外のシステムに設けられていてもよい。即ち、当該処理はランキングスコア等によるソートのような周知の処理である。例えば、図9

(A)に示す回答61又は62をORDERBY部に従ってソートすると、図9(B)に示すような結果を得る。即ち、文書名「0323.txt」におけるノード番号\$8に対応する株価が最も高いという結果を得る。

【0058】最終回答構成処理部7は、（ソート等された）回答61を第1最終回答構成処理部71において処理するか、又は、（ソート等された）回答62を第2最終回答構成処理部72において処理する。例えば、「ソートの結果の上位1件のみを出力せよ」という指示（図8(A)には図示せず）であれば、図9(C)に示すように、上述のAA社の最高の株価を出力する。

【0059】図10は、構造化文書格納検索システムが

実行する構造化文書処理フローを示す。

【0060】実行すべき処理が格納処理であるか否かを調べる（ステップS1）。

【0061】格納処理である場合、格納処理を実行して（ステップS2）、処理を終了する。これについては、図11を参照して後述する。

【0062】格納処理でない場合、実行すべき処理が検索回答処理であるか否かを調べる（ステップS3）。

【0063】検索回答処理である場合、検索回答処理を実行して（ステップS4）、処理を終了する。これについては、図12を参照して後述する。

【0064】検索回答処理でない場合、処理を終了する。

【0065】図11は、構造化文書格納処理部2が実行する構造化文書格納処理フローを示す。

【0066】構造化文書格納処理部2が格納する構造化文書の原文1を取得する（ステップS11）。

【0067】構造化文書格納処理部2の原文格納処理部25が、原文1を格納するか否かを調べる（ステップS12）。格納しない場合、ステップS14に進む。

【0068】格納する場合、原文格納処理部25が原文格納処理を行うことにより、当該原文1を原文格納部9に格納する（ステップS13）。

【0069】文書分解処理部21が文書分解処理により当該原文1を分解し、構造構成処理部22が構造構成処理によりノード番号配列、タグリスト32、木を構成し、これらを構造格納部3に格納する（ステップS14）。

【0070】図12は、構造化質問処理部5及び最終回答構成処理部7が実行する構造化質問処理及び回答処理フローを示す。

【0071】検索処理部51が、構造化文書の質問4を受け取る（ステップS21）。

【0072】検索処理部51が、検索処理により該当文書を得る（ステップS22）。これについては、図13を参照して後述する。この図13に示す処理は、各文書（1A乃至1C）毎即ち、タグリスト32（32A乃至32C）毎に繰り返し行われる。これにより、質問4に該当する文書（の文書名）のみが得られる。

【0073】ノード番号回答処理部52が、ノード番号回答処理により回答格納部6に一時結果を格納する（ステップS23）。これについては、図14乃至図16を参照して後述する。

【0074】最終回答構成処理部7が、回答格納部6に格納された一時結果に基づいて、最終的な回答を構成して出力する（ステップS24）。

【0075】図13は、検索処理部51が実行する検索処理フローを示す。即ち、検索処理によりタグ番号からノードを取得する処理である。

【0076】ノード返却集合Rを空にする（ステップS

31)。

【0077】タグリスト構造にタグ番号をキーにアクセスしたタグリストポイントpを取得する(ステップS32)。

【0078】タグリストポイントpが未設定であるか否かを調べる(ステップS33)。未設定である場合、ステップS36に進む。

【0079】未設定でない場合、当該ポイントpの指すノードnをノード返却集合Rに追加する(ステップS34)。

【0080】当該ポイントpの次のタグリストポイントを、新しいタグリストポイントpとして設定し(ステップS35)、ステップS33以下を繰り返す。

【0081】以上により得たノード返却集合Rを、取得したノードとして返却し(ステップS36)、処理を終了する。例えば、図5(B)に示すように、タグ<株価>からそのタグ番号を知り、これを用いてポイントP1及びP2の指すノード\$8及び\$11をノード返却集合Rに得る。ノード返却集合Rが空である文書1(d)は質問4に該当しない文書である。

【0082】図14は、ノード番号回答処理部52及び最終回答構成処理部7が実行するノード番号回答処理及び最終回答構成処理フローを示す。

【0083】ノード番号回答処理部52が、検索質問4を満たす結果を全て抽出して、これを一時結果として回答格納部6に蓄積する(ステップS41)。これについては、図15を参照して後述する。

【0084】最終回答構成処理部7が、必要に応じて、回答格納部6に蓄積された一時結果を、前述のように、ランキングスコアによるソート等の操作をする(ステップS42)。この処理は、後述するように、必要に応じて実行される。

【0085】最終回答構成処理部7が、操作の結果に基づいて、最終回答を生成する(ステップS43)。これについては、図16を参照して後述する。

【0086】例えば、上位1000件のランキング検索を行なう場合、ステップS41で検索質問4を満たす結果を全て蓄積し、ステップS42でランキングスコアによるソートを行ない、ステップS43で上位1000件について表示可能な回答を生成する。

【0087】なお、ステップS42は、必要に応じて、適宜実行される。即ち、上述の例において、ステップS41において検索された結果である要素の個数が1000件未満である場合、当該結果を全て出力できるので、ランキングスコアによるソートは必要ない場合もある。この場合、ステップS42の実行は省略される。

【0088】図15は、ノード番号回答処理部52が実行する一時結果の蓄積処理フローを示す。

【0089】前処理を行う(ステップS51)。即ち、検索を行ない、検索質問4にマッチする文書を文書集合

Dにセットし、検索質問4の回答指示を要素毎に分解して回答指示要素集合Sにセットし、結果格納集合Rを空にする。

【0090】文書集合Dが空か又は回答指示要素集合Sが空か否かを調べる(ステップS52)。当該集合D又はSが空の場合、ステップS60へ進む。

【0091】当該集合D及びSが空でない場合、 $(d, s) \in (D, S)$ である (d, s) を1個取り出す(ステップS53)。

【0092】文書d内で回答指示sにマッチする分岐なし部分木の最下部の(最も葉に近い)ノードの通し番号を $Nn(d, s)$ にセットする(ステップS54)。

【0093】 $Nn(d, s)$ が空でないならば、文書dのポイントがdp、回答指示要素sのポイントがspのとき、 $r(dp, sp, Nn(d, s))$ を結果格納集合Rに追加する(ステップS55)。

【0094】結果格納集合Rに結果rが複数個存在する場合は、1つにする(ステップS56)。

【0095】 $(d, s) \in (D, S)$ である全ての (d, s) について処理したか否かを調べる(ステップS57)。処理していない場合、ステップS53以下を繰り返す。

【0096】処理した場合、結果格納集合Rが空か否かを調べる(ステップS58)。空の場合、ステップS60へ進む。

【0097】空でない場合、回答指示要素集合Sと結果格納集合Rを一時結果として格納して、処理を終了する(ステップS59)。

【0098】結果格納集合Rを一時結果として格納する(ステップS60)。例えば、回答指示要素sp=<株価>について、図3乃至図5に示すdp=「0323.txt」なる文書のノード番号 $Nn(d, s) = \$8$ を、結果格納集合Rに得る。即ち、 $(0323.txt, <株価>, \$8)$ を得る(図19においても、略同様)。

【0099】図16及び図17は、両者で1つのフローを構成し、最終回答構成処理部7が実行する最終回答構成処理フローを示す。

【0100】前処理を行う(ステップS61)。即ち、最終結果格納集合Rを空にする。ステップS41(又はS42)により格納した回答指示要素集合をSt、一時結果格納集合をRtとする。

【0101】集合Stが空又は集合Rtが空か否かを調べる(ステップS62)。空の場合、ステップS67へ進む。

【0102】空でない場合、 $rt(dp, sp, Nn(d, s)) \in Rt$ である $rt(dp, sp, Nn(d, s))$ を1個取り出す(ステップS63)。

【0103】文書dpからノード番号配列(34、図6参照)を用いてノード通し番号 $Nn(d, s)$ のノードnを取得する(ステップS64)。

【0104】回答指示要素集合 S_t からポインタ s_p の回答指示 s を取得する(ステップS65)。

【0105】木集合 T を空にする(ステップS66)。

【0106】回答指示 s が単一タグか否かを調べる(ステップS67)。単一タグである(パス又はパスワイルドカードでない)場合、ステップS71へ進む。ここで、単一タグとは、例えば図4において、ノード「株価」に要素「4, 020」が繋がっているようなタグを言い、単一タグでないとは、例えば図4において、ノード「会社情報」にノード「会社名」及び要素「AA社」が繋がっているようなタグを言う。

【0107】単一タグでない(パス又はパスワイルドカードである)場合、回答指示 s の分岐なし部分木の最上ノード(最もルートに近いノード)のタグ名 t_n を取得する(ステップS68)。

【0108】木構造情報を用いてノード n からルートノードまで親方向へたどる(ステップS69)。

【0109】当該たどる過程において、タグ名 t_n と一致するノード n_t があったら、その各々について、ノード n_t とノード n の間の分岐なし部分木を、木集合 T に追加する(ステップS70)。

【0110】ノード n に子(子のノード)がある場合、木構造情報を用いてそれ以下の全てのノードを取得しノード集合 N_c とする(ステップS71)。

【0111】 $t \in T$ である木 t を1個取り出す(ステップS72)。

【0112】ノード集合 N を空にする(ステップS73)。

【0113】木構造 t を構成するノードをノード集合 N に追加する(ステップS74)。

【0114】全てのノード集合 N_c の要素を N に追加する(ステップS75)。

【0115】結果 $r(d_p, s_p, N)$ を最終結果格納集合 R に追加する(ステップS76)。

【0116】 $t \in T$ である全ての木 t について処理したか否かを調べる(ステップS77)。全ての木 t について処理していない場合、ステップS72以下を繰り返す。

【0117】全ての木 t について処理した場合、 $r_t(d_p, s_p, N_n(d, s)) \in R_t$ である全ての $r_t(d_p, s_p, N_n(d, s))$ について処理したか否かを調べる(ステップS78)。全ての木 t について処理していない場合、ステップS63以下を繰り返す。

【0118】全ての木 t について処理した場合、最終結果格納集合 R を結果とする(ステップS79)。例えば、前述の(0323.txt, <株価>, \$8)に基づいて、文書0323.txtのノード\$8の内容「株価」をノード(配列)34から得て、これにその要素%5「4, 020」を追加して N とし、結果 $r(0323.txt, <株価>, 株価及び4, 020)$ を集合 R に得る(図20においても、

略同様)。

【0119】以上に述べた図14に示すノード番号回答処理及び最終回答構成処理に代えて、以下の図18に示すように、ノード番号回答処理及び最終回答構成処理を実行してもよい。図18は、ノード番号回答処理部52及び最終回答構成処理部7が実行するノード番号回答処理及び最終回答構成処理フローを示す。

【0120】ノード番号回答処理部52が、ステップS41と同様に、検索質問4を満たす結果を全て抽出して、これを一時結果として回答格納部6に蓄積する(ステップS81)。これについては、図19を参照して後述する。

【0121】最終回答構成処理部7が、ステップS42と同様に、必要に応じて、回答格納部6に蓄積された一時結果を、前述のように、ランキングスコアによるソート等の操作をする(ステップS82)。この処理は、前述のステップS42と同様に、必要に応じて実行される。

【0122】最終回答構成処理部7が、ステップS43と同様に、操作の結果に基づいて、最終回答を生成する(ステップS83)。これについては、図20を参照して後述する。

【0123】図19は、ノード番号回答処理部52が実行する一時結果の蓄積処理フローを示す。

【0124】前処理を行う(ステップS91)。即ち、検索を行ない、検索質問4にマッチする文書を文書集合 D にセットし、検索質問4の回答指示を要素毎に分解して回答指示要素集合 S にセットし、結果格納集合 R を空にする。

【0125】文書集合 D が空か又は回答指示要素集合 S が空か否かを調べる(ステップS92)。当該集合 D 又は S が空の場合、ステップS99へ進む。

【0126】当該集合 D 及び S が空でない場合、 $(d, s) \in (D, S)$ である (d, s) を1個取り出す(ステップS93)。

【0127】文書 d 内で回答指示 s にマッチする分岐なし部分木の最上部の(最もルートに近い)ノードの通し番号を $N_r(d, s)$ にセットし、最下部の(最も葉に近い)ノードの通し番号を $N_l(d, s)$ にセットする(ステップS94)。

【0128】 $N_r(d, s)$ 及び $N_l(d, s)$ が空でないならば、文書 d のポインタが d_p 、回答指示要素 s のポインタが s_p のとき、 $r(d_p, s_p, N_r(d, s), N_l(d, s))$ を結果格納集合 R に追加する(ステップS95)。

【0129】 $(d, s) \in (D, S)$ である全ての (d, s) について処理したか否かを調べる(ステップS96)。処理していない場合、ステップS93以下を繰り返す。

【0130】処理した場合、結果格納集合 R が空か否か

を調べる(ステップS97)。空の場合、ステップS99へ進む。

【0131】空でない場合、回答指示要素集合Sと結果格納集合Rを一時結果として格納して、処理を終了する(ステップS98)。

【0132】結果格納集合Rを一時結果として格納する(ステップS99)。

【0133】図20は、最終回答構成処理部7が実行する最終回答構成処理フローを示す。

【0134】前処理を行う(ステップS101)。即ち、最終結果格納集合Rを空にする。ステップS81(又はS82)により格納した回答指示要素集合をSt、一時結果格納集合をRtとする。

【0135】集合Stが空又は集合Rtが空か否かを調べる(ステップS102)。空の場合、ステップS110へ進む。

【0136】空でない場合、 $rt(dp, sp, Nr(d, s), Nl(d, s)) \in Rt$ である $rt(dp, sp, Nr(d, s), Nl(d, s))$ を1個取り出す(ステップS103)。

【0137】ノード集合Nを空にする(ステップS104)。

【0138】文書dpからノード番号配列を用いてノード通し番号Nr(d, s)のノードnr、Nl(d, s)のノードnlを取得する(ステップS105)。

【0139】木構造情報を用いてノードnlからノードnrまでたどり、それらのノードを全てノード集合Nに追加する(ステップS106)。

【0140】ノードnlに子(子のノード)がある場合は木構造情報を用いてそれ以下の全てのノードを取得しノード集合Nに追加する(ステップS107)。

【0141】結果 $r(dp, sp, N)$ を最終結果格納集合Rに追加する(ステップS108)。

【0142】 $rt(dp, sp, Nr(d, s), Nl(d, s)) \in Rt$ である全ての $rt(dp, sp, Nr(d, s), Nl(d, s))$ について処理したか否かを調べる(ステップS109)。処理していない場合、ステップS103以下を繰り返す。

【0143】処理した場合、最終結果格納集合Rを結果として終了する(ステップS110)。

【0144】なお、図14乃至図17と図18乃至図20との対比から判るように、これらの処理では、回答時の処理が少し異なる。回答処理の一時格納においては、前者は回答指示の葉の部分のみ蓄積するが、後者は回答指示のルートの部分も蓄積する。従って、一時蓄積する回答の分量は前者の方が少なくて済む。一方、回答処理の最終回答生成においては、前者は回答指示のルートに相当するノードを検索し直すが、後者はその処理を行なう必要がない。従って、最終回答生成に要する処理の時間は後者のほうが短かくて済む。

【0145】

【発明の効果】以上説明したように、本発明によれば、構造化文書格納装置において、構造化文書の原文に対応する木構造に基づいて生成されたタグ番号配列とタグリストとからなるタグリスト構造を備えることにより、構造化文書の高速度での検索に適したデータ構造を得ることができる。

【0146】また、本発明によれば、構造化文書検索装置において、木構造に基づいて生成されたタグ番号配列とタグリストとからなるタグリスト構造を利用して検索を行うことにより、検索対象となる要素に至るまで順に木構造をたどることなく、かつ、検索対象となる要素以外の要素を全く検索することなく、検索対象となる要素を極めて高速に、検索対象となる要素を検索することができる。

【0147】また、本発明によれば、構造化文書検索装置において、質問に対する回答をノード番号により構成することにより、検索処理の結果として当該結果である要素の実体を複製して蓄積する必要がなく、少ないメモリ資源で当該結果を蓄積することができるので、検索結果を一時的に大量に蓄積して、これに対して操作を行なって選別して出力するような処理を、容易に行うことができる。

【図面の簡単な説明】

【図1】本発明の原理構成図である。

【図2】構造化文書格納検索システム構成図である。

【図3】原文説明図である。

【図4】木構造説明図である。

【図5】タグリスト構造説明図である。

【図6】構造化文書格納検索説明図である。

【図7】構造化文書格納検索説明図である。

【図8】構造化文書格納検索説明図である。

【図9】構造化文書格納検索説明図である。

【図10】構造化文書処理フローを示す。

【図11】構造化文書格納処理フローを示す。

【図12】構造化質問処理及び回答処理フローを示す。

【図13】検索処理フローを示す。

【図14】ノード番号回答処理及び最終回答構成処理フローを示す。

【図15】一時結果の蓄積処理フローを示す。

【図16】最終回答構成処理フローを示す。

【図17】最終回答構成処理フローを示す。

【図18】ノード番号回答処理及び最終回答構成処理フローを示す。

【図19】一時結果の蓄積処理フローを示す。

【図20】最終回答構成処理フローを示す。

【符号の説明】

3 構造格納部

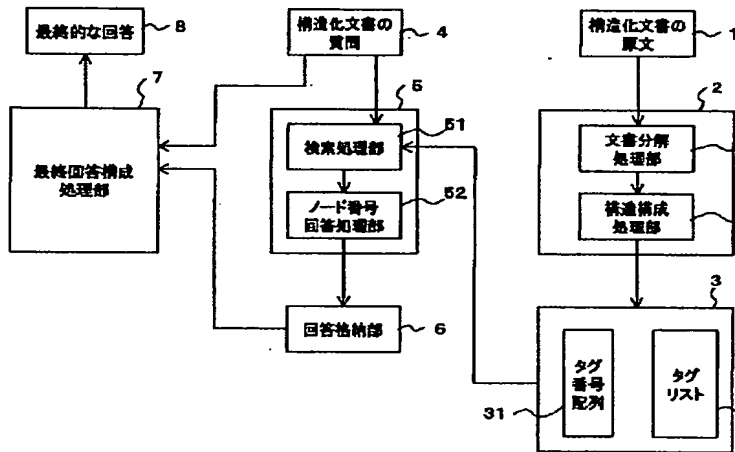
6 回答格納部

21 文書分解処理部

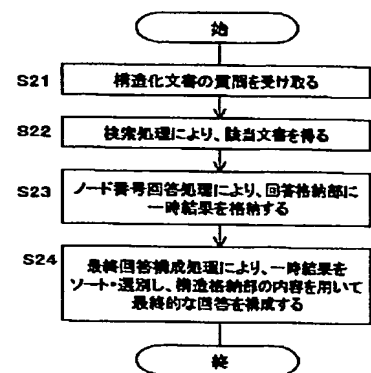
2 2 構造構成処理部
2 3 ノード番号付与処理部
2 4 タグリスト構成処理部
3 1 タグ番号配列
3 2 タグリスト構成 (群)

5 1 検索処理部
5 2 ノード番号回答処理部
7 1 第1最終回答処理部
7 2 第2最終回答処理部

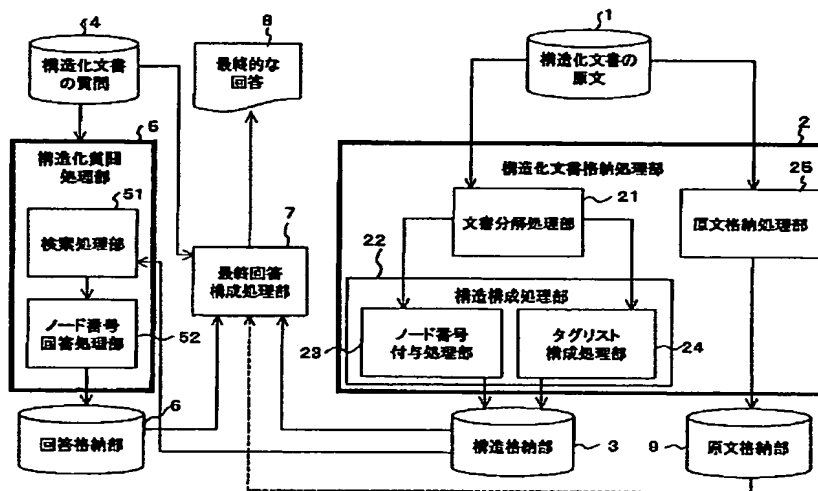
【図 1】



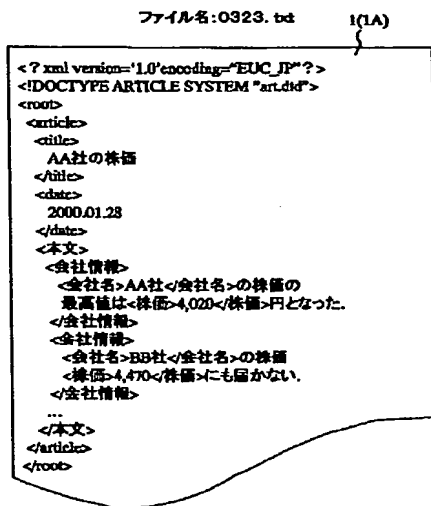
【図 1 2】



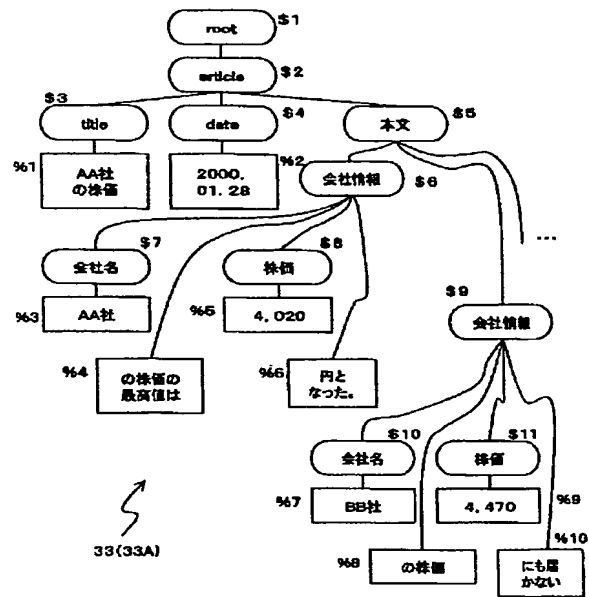
【図 2】



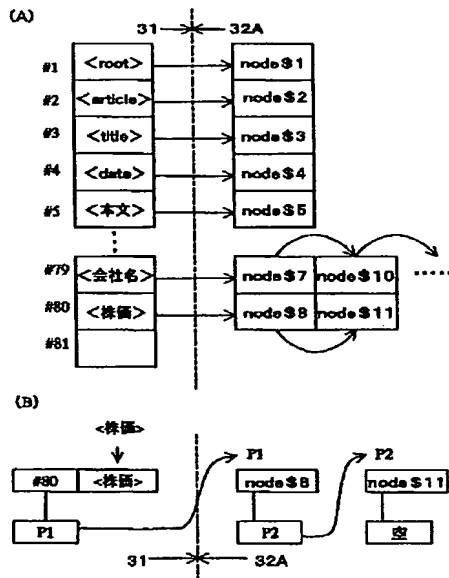
【図3】



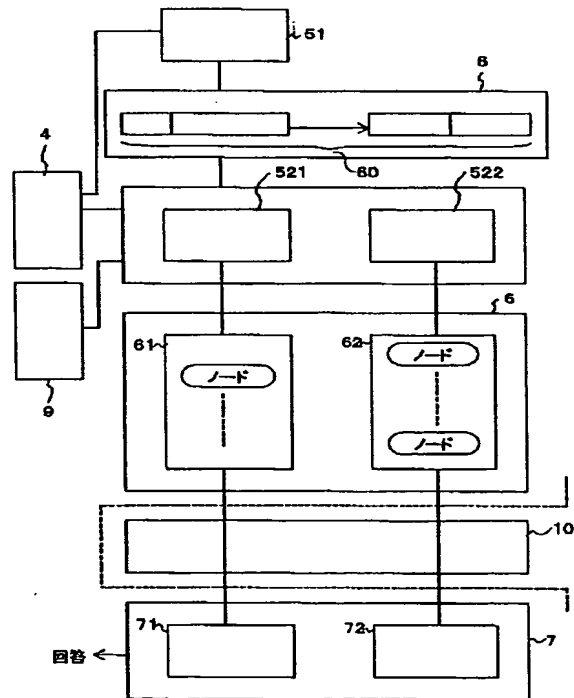
【図4】



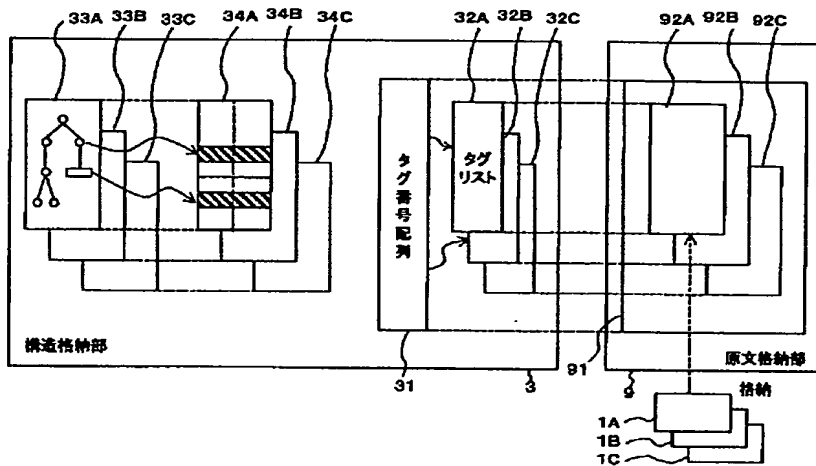
【図5】



【図7】



【図6】



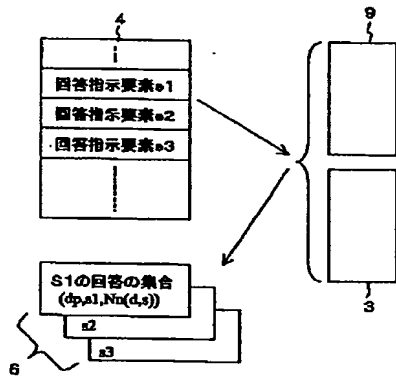
【図8】

【図9】

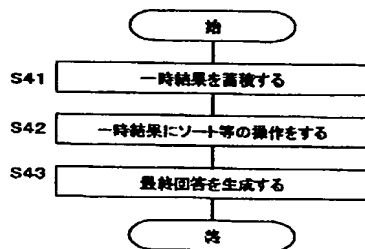
(A)

```
SELECT 会社情報#L/株価
WHERE <会社情報#><会社名>AA社</>
ORDERBY 会社情報#L/株価
```

(B)



【図14】



(A)

文書名	回答指示要素	ノード通し番号	ソート値
0001	株価	8	3, 670
0197	株価	8	3, 090
0224	株価	14	3, 820
0323	株価	8	4, 020
0457	株価	22	3, 800
0598	株価	57	2, 380

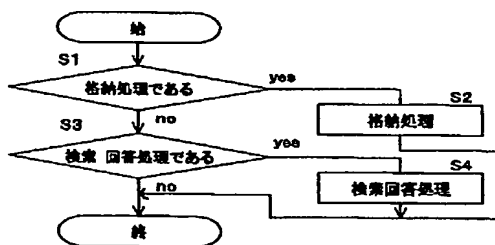
(B)

文書名	回答指示要素	ノード通し番号	ソート値
0323	株価	8	4, 020
0224	株価	14	3, 820
0457	株価	22	3, 800
0001	株価	8	3, 670
0197	株価	8	3, 090
0598	株価	57	2, 380

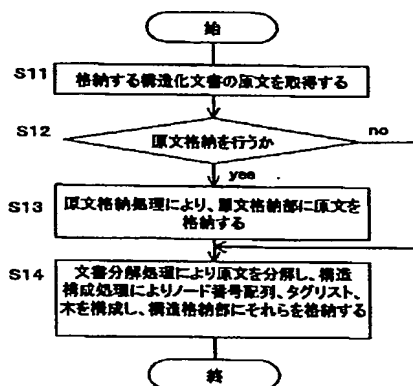
(C)

<株価>4, 020</株価>

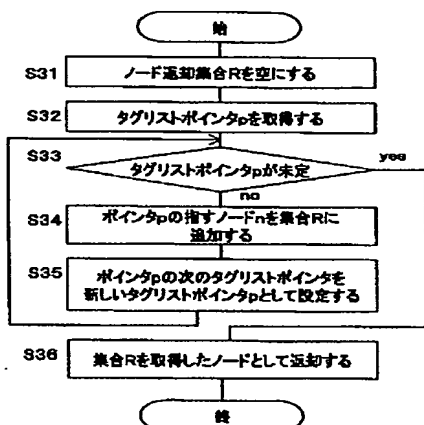
【図10】



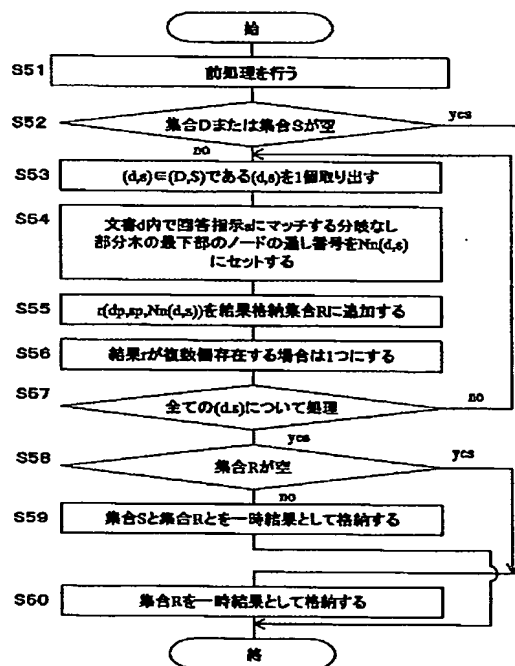
【図11】



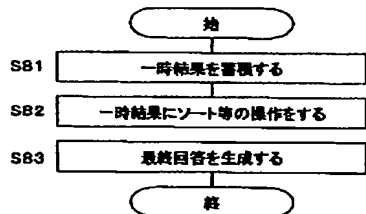
【図13】



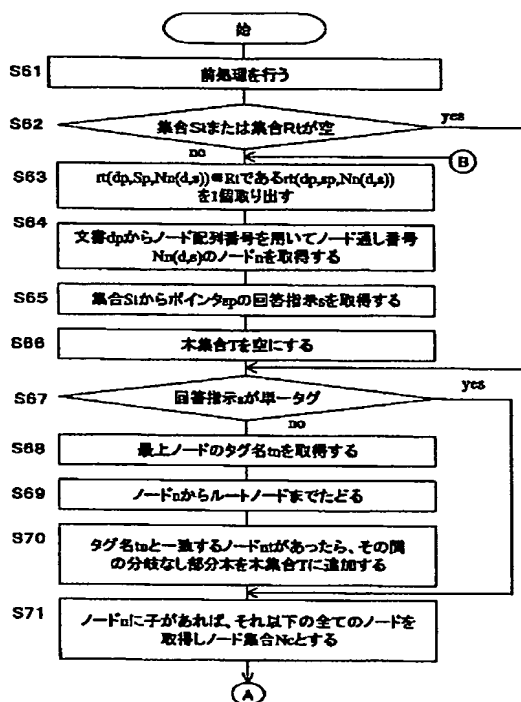
【図15】



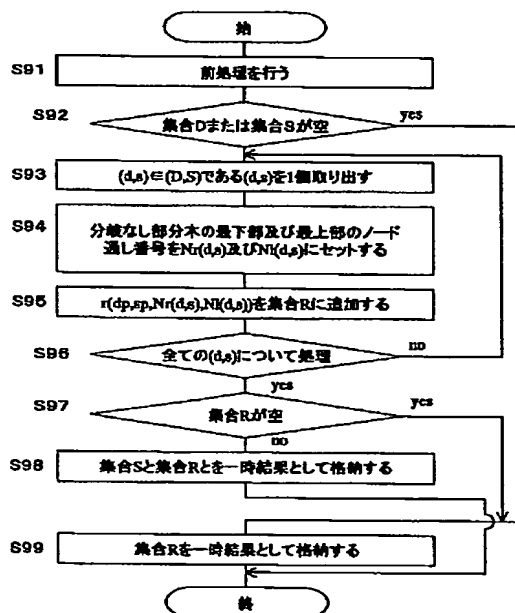
【図18】



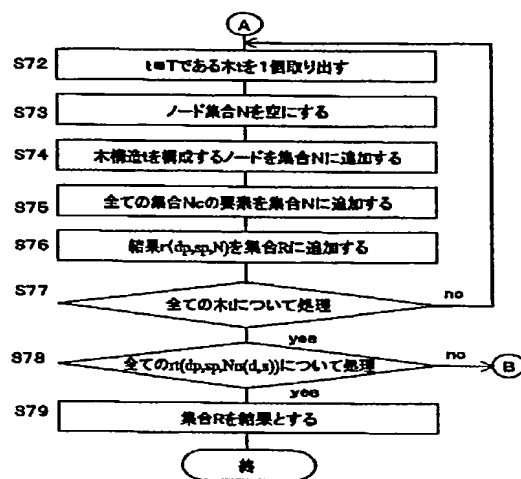
【図16】



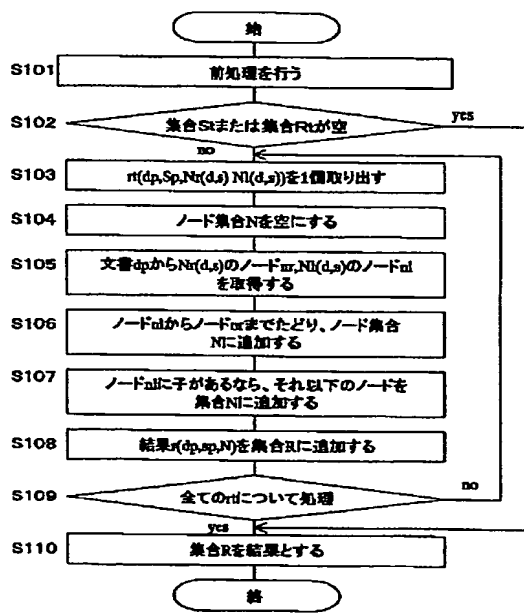
【図19】



【図17】



【図20】



This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record.

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☒ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.